Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет
по лабораторной работе №1

«Разведочный анализ данных. Исследование и визуализация данных»

по курсу «Технологии машинного обучения»

ИСПОЛНИТЕЛЬ:

Голубкова С.В.
Группа РТ5-61Б

_____

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

_____

Москва 2020

Лабораторная работа № 1

1. Цель лабораторной работы

Изучение различных методов визуализация данных.

2. Задание

- Выбрать набор данных (датасет).
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например, из Scikit-learn.
- Для лабораторных работ не рекомендуется выбирать датасеты большого размера.
- Создать ноутбук, который содержит следующие разделы:
    1. Текстовое описание выбранного Вами набора данных.
    2. Основные характеристики датасета.
    3. Визуальное исследование датасета.
    4. Информация о корреляции признаков.
- Сформировать отчет и разместить его в своем репозитории на github.

3. Выполнение работы

Лабороторная работа №1

1) Текстовое описание набора данных

Scikit-learn поставляется с несколькими небольшими стандартными наборами данных, которые не требуют загрузки файла с какого-либо внешнего веб-сайта.

Был выбран dataset для распознавания вина (*wine recognition dataset*). Эти данные являются результатами химического анализа вин, выращенных в одном и том же регионе в Италии и представленными тремя различными культиваторами. Существуют тринадцать различных измерений, проведенных для различных компонентов, найденных в трех типах вина.

Имеются следующие колонки:

- Alcohol - крепость алкогольных напитков обозначают в процентах от объема
- Malic acid - содержание яблочной кислоты в вине
- Ash - содержание золы (мг/л)
- Alcalinity of ash - щелочность золы в граммах на литр карбоната калия
- Magnesium - содержание магния в вине в мг/л
- Total phenols - суммарное содержание фенолов
- Flavanoids - концентрация полифенолов (г/литр)
- Nonflavanoid phenols - концентрация нефлавоноидов
- Proanthocyanins - концентрация антоцианов (г/литр)
- Color intensity - интенсивность цвета
- Hue - оттенок
- OD280/OD315 of diluted wines - разбавленность вина
- Proline - содержание пролина (аминокислоты) в вине (мг/литр)
- Target - целевой признак: класс 0, класс 1, класс 2

Преобразование датасета Scikit-learn в Pandas Dataframe

In [2]:
```
import numpy as np
import pandas as pd
from sklearn.datasets import *
```

In [3]:
```
wine = load_wine()
```

In [4]:
```
type(wine)
```

Out[4]:
```
sklearn.utils.Bunch
```

In [5]:
```
for x in wine:
    print(x)
```

```
data
target
target_names
DESCR
feature_names
```

In [6]:
```
wine['target_names']
```

Out[6]:
```
array(['class_0', 'class_1', 'class_2'], dtype='<U7')
```

In [7]:
```
wine['feature_names']
```

Out[7]:
```
['alcohol',
 'malic_acid',
 'ash',
 'alcalinity_of_ash',
 'magnesium',
 'total_phenols',
 'flavanoids',
 'nonflavanoid_phenols',
 'proanthocyanins',
 'color_intensity',
 'hue',
 'od280/od315_of_diluted_wines',
 'proline']
```

In [8]:
```
# Размерность данных
wine['data'].shape
```

Out[8]:
```
(178, 13)
```

In [9]:
```
wine['target'].shape
```

Out[9]:
```
(178,)
```

In [10]:
```
#Преобразование в Pandas DataFrame
data1 = pd.DataFrame(data= np.c_[wine['data'], wine['target']],
                     columns= wine['feature_names'] + ['target'])
```

In [11]:
```
data1
```

Out[11]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	
...	...	...	...	...	...	...	...	...	...	...	...	...
173	13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	7.70	0.64	
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	0.43	1.41	7.30	0.70	
175	13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	10.20	0.59	
176	13.17	2.59	2.37	20.0	120.0	1.65	0.68	0.53	1.46	9.30	0.60	
177	14.13	4.10	2.74	24.5	96.0	2.05	0.76	0.56	1.35	9.20	0.61	

178 rows × 14 columns

2) Основные характеристики датасета

In [12]:
```
# Первые 5 строк
data1.head()
```

Out[12]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od3
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	

In [13]:
```
# Размер датасета - 178 строк, 14 колонок
data1.shape
```

Out[13]:
```
(178, 14)
```

In [14]:
```
total_count = data1.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 178
```

In [15]:
```
data1.columns
```

Out[15]:
```
Index(['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium',
       'total_phenols', 'flavanoids', 'nonflavanoid_phenols',
       'proanthocyanins', 'color_intensity', 'hue',
       'od280/od315_of_diluted_wines', 'proline', 'target'],
      dtype='object')
```

In [16]:
```
# Список колонок с типами данных
data1.dtypes
```

Out[16]:
```
alcohol                         float64
malic_acid                      float64
ash                             float64
alcalinity_of_ash               float64
magnesium                       float64
total_phenols                   float64
flavanoids                      float64
nonflavanoid_phenols            float64
proanthocyanins                 float64
color_intensity                 float64
hue                             float64
od280/od315_of_diluted_wines    float64
proline                         float64
target                          float64
dtype: object
```

In [17]:
```
# Нет пустых значений
for col in data1.columns:
    temp_null_count = data1[data1[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
alcohol - 0
malic_acid - 0
ash - 0
alcalinity_of_ash - 0
magnesium - 0
total_phenols - 0
flavanoids - 0
nonflavanoid_phenols - 0
proanthocyanins - 0
color_intensity - 0
hue - 0
od280/od315_of_diluted_wines - 0
proline - 0
target - 0
```

In [18]:
```
# Основные статистические характеристки набора данных
data1.describe()
```

Out[18]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	1
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.058090	
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318286	
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280000	
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220000	
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690000	
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200000	
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.000000	

In [19]:
```
# Уникальные значения для целевого признака
data1['target'].unique()
```

Out[19]:
```
array([0., 1., 2.])
```

3) Визуальное исследование датасета

Подключение библиотек:

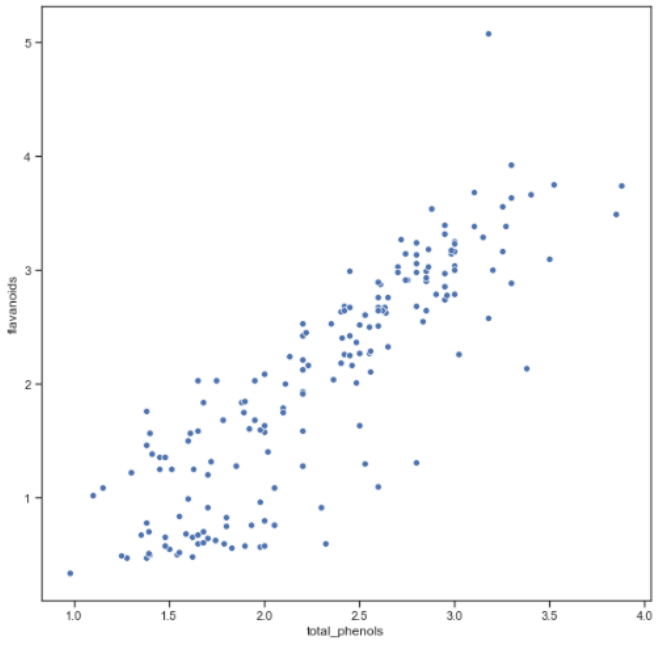
In [20]:
```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Диаграмма рассеяния

Поможет определить имеется ли линейная зависимость между колонками 'total_phenols' и 'flavonoids'.

In [21]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='total_phenols', y='flavanoids', data=data1)
```

Out[21]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xd7ddc70>
```

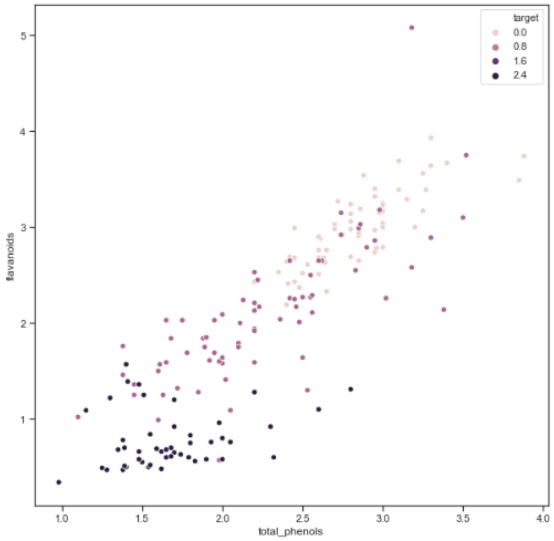


Можно видеть что между этими полями имеет место зависимость, близкая к линейной.

Посмотрим насколько на эту зависимость влияет целевой признак.

In [22]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='total_phenols', y='flavanoids', data=data1, hue='target')
```

Out[22]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xec4b530>
```

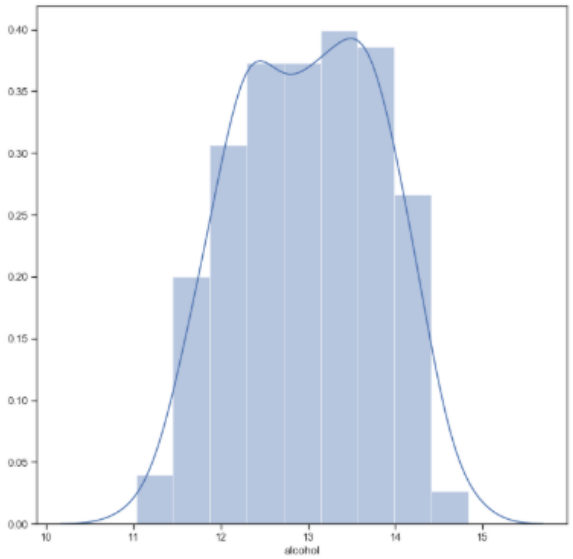


Гистограмма

Позволяет оценить плотность вероятности распределения данных.

In [23]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data1['alcohol'])
```

Out[23]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xe9e2190>
```

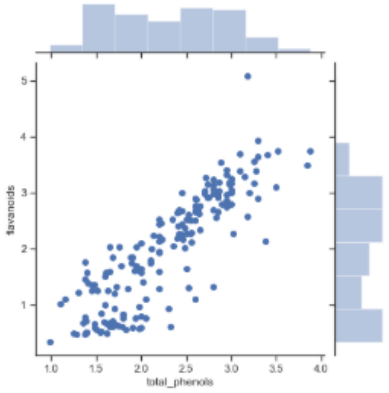


Jointplot

Комбинация гистограмм и диаграмм рассеивания.

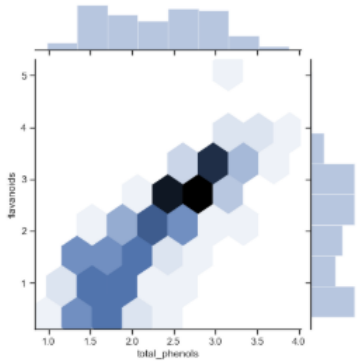
In [24]:
```
sns.jointplot(x='total_phenols', y='flavanoids', data=data1)
```

Out[24]:
```
<seaborn.axisgrid.JointGrid at 0xd2bb510>
```



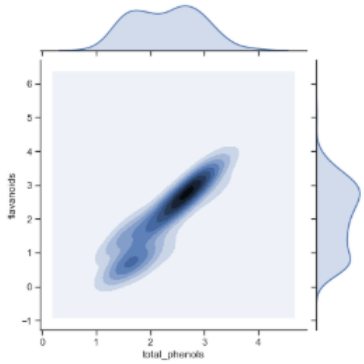
In [25]:
```
sns.jointplot(x='total_phenols', y='flavanoids', data=data1, kind="hex")
```

Out[25]:
```
<seaborn.axisgrid.JointGrid at 0xee3e5d0>
```



In [26]:
```
sns.jointplot(x='total_phenols', y='flavanoids', data=data1, kind="kde")
```

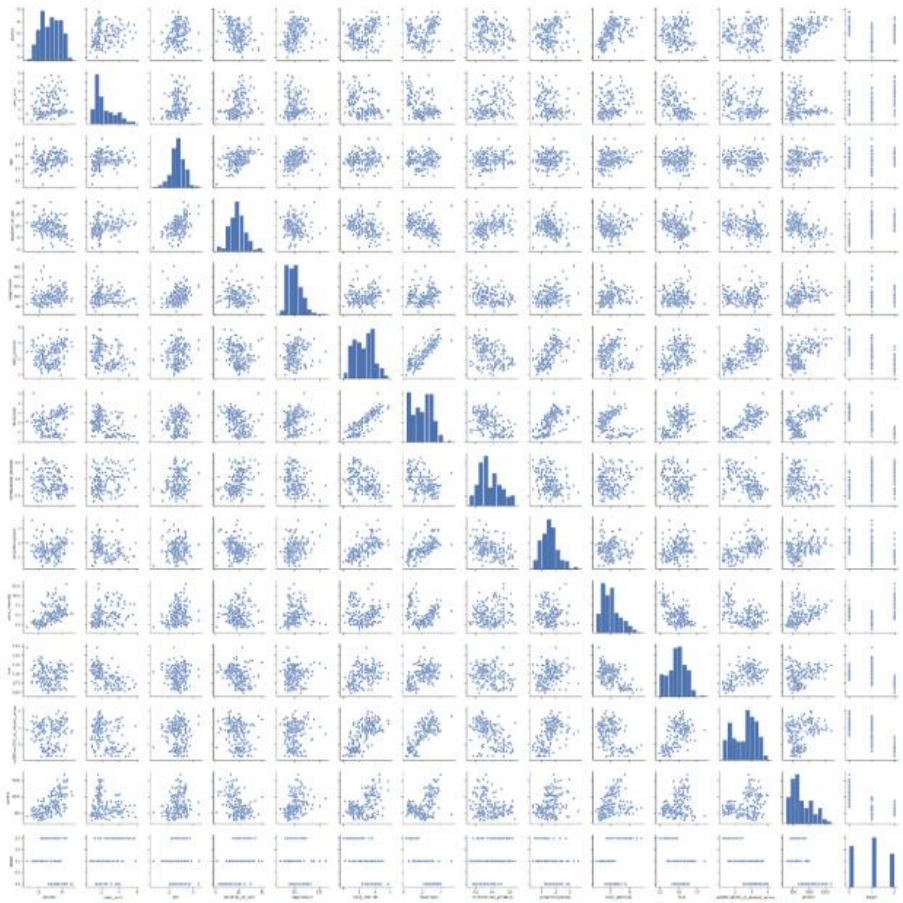
Out[26]:
```
<seaborn.axisgrid.JointGrid at 0xee211d0>
```



Парные диаграммы

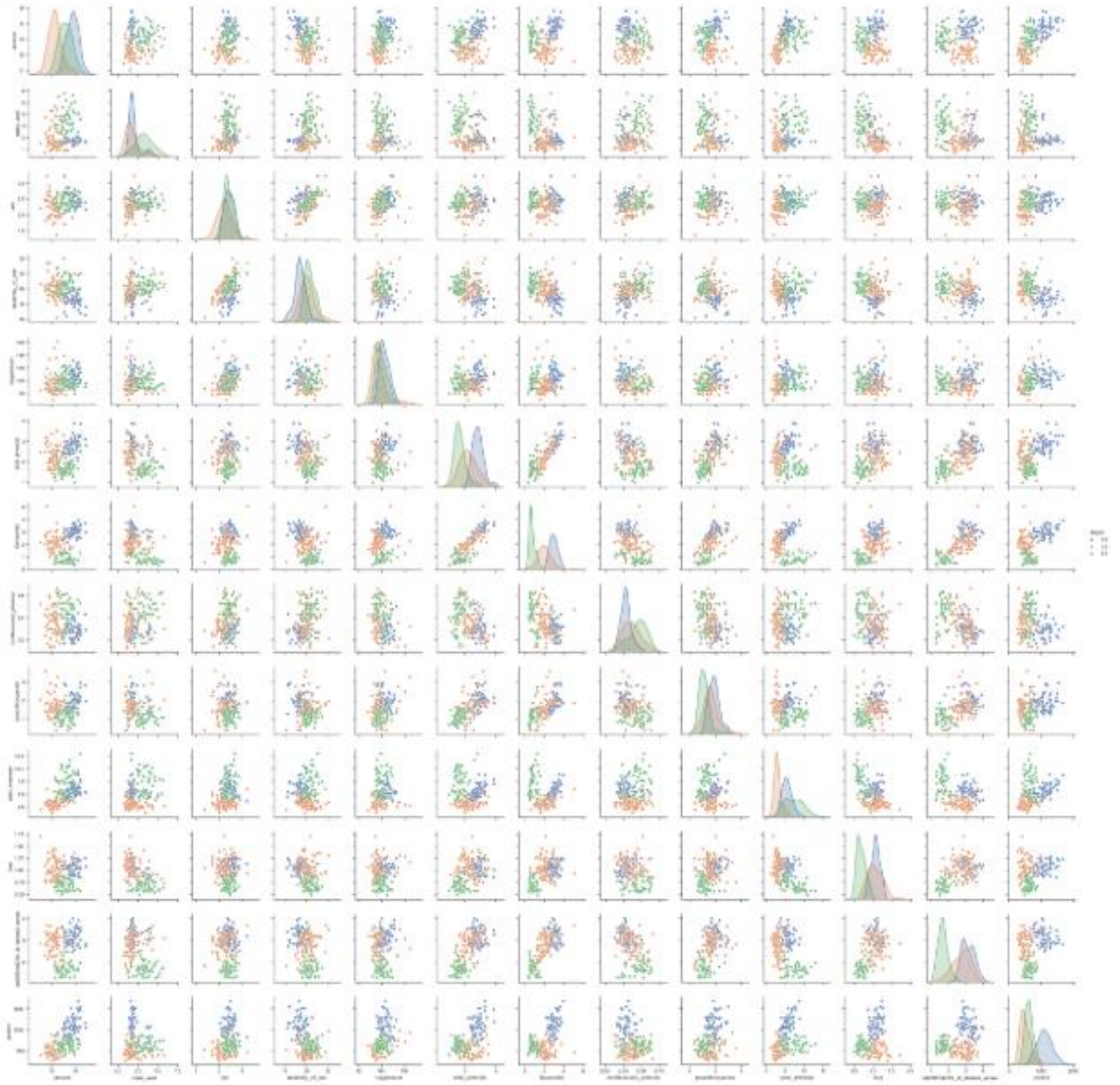
In [27]:
```
sns.pairplot(data1)
```

Out[27]:
```
<seaborn.axisgrid.PairGrid at 0xd052710>
```



In [28]:
```
sns.pairplot(data1, hue="target")
```

Out[28]:
```
<seaborn.axisgrid.PairGrid at 0xd8ceb30>
```

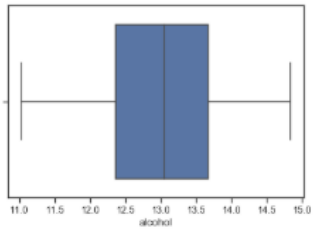


Ящик с усами

Отображает одномерное распределение вероятности параметра 'alcohol'.

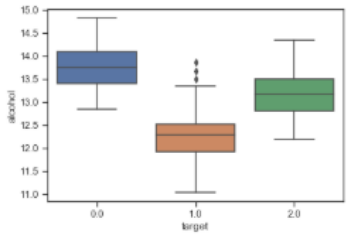
In [29]:
```
sns.boxplot(x=data1['alcohol'])
```

Out[29]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a6a3910>
```



In [30]:
```
# Распределение параметра Alcohol сгруппированные по Target
sns.boxplot(x=data1['target'], y=data1['alcohol'])
```

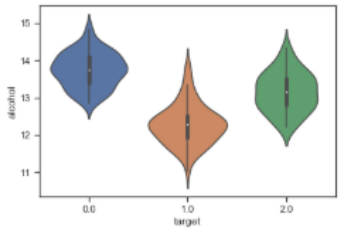
Out[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1be1fbf0>
```



Violin plot

In [31]:
```
# Распределение параметра Alcohol сгруппированные по Target
sns.violinplot(x='target', y='alcohol', data=data1)
```

Out[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1beb7430>
```



4) Информация о корреляции признаков

In [32]:
```
data1.corr()
```

Out[32]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.136698
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0.220746
ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.009652
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0.197327
magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0.236441
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.612413
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.652692
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0.365845
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.365845	1.000000
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.139057	-0.025250
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.262640	0.295544
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.503270	0.519067
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.311385	0.330417
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0.499130

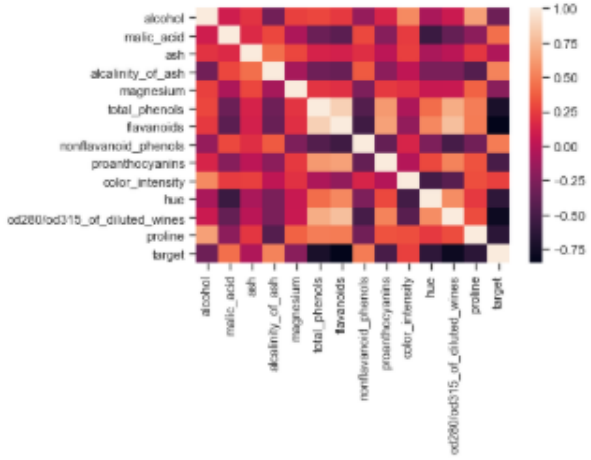
На основе корреляционной матрицы можно сделать следующие выводы (значения корреляции взяты по модулю):

- Целевой признак наиболее сильно коррелирует с общим колличеством фенолов (0.72), разбавленностью вина (0.79) и концентрацией полифенолов (0.85). Эти признаки обязательно следует оставить в модели.
- Целевой признак отчасти коррелирует с концентрацией яблочной кислоты (0.43), щелочностью золы (0.52), концентрацией нефлавоноидов (0.49), концентрацией антоцианов (0.5), оттеноком вина(0.62) и концентарцией пролина (0.63). Этоти признаки стоит также оставить в модели.
- Целевой признак слабо коррелирует с концентрацией золы (0.04), alcohol (0.33), интенсивностью цвета вина (0.27) и содержанием магния в вине (0.21). Скорее всего эти признаки стоит исключить из модели, возможно они только ухудшат ее качество.
- Общая концентрация фенолов и концентрация полифенолов очень сильно коррелируют между собой (0.86). Это неудивительно, ведь флавоноиды – обширный класс низкомолекулярных многоатомных фенолов растительного происхождения.
- Также можно сделать вывод, что выбирая из признаков total_phenols и flavanoids лучше выбрать flavanoids, потому что он сильнее коррелирован с целевым признаком. Если линейно зависимые признаки сильно коррелированы с целевым, то оставляют именно тот признак, который коррелирован с целевым сильнее.

Тепловые карты

In [33]:
```
sns.heatmap(data1.corr())
```

Out[33]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a8c7c50>
```



In [34]:
```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(40,10))
sns.heatmap(data1.corr(method='pearson'),cmap='YlGnBu', ax=ax[0])
sns.heatmap(data1.corr(method='kendall'), cmap='YlGnBu', ax=ax[1])
sns.heatmap(data1.corr(method='spearman'), cmap='YlGnBu', ax=ax[2])
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```